

An investigation of item bias in the four-tier diagnostic test using Rasch model

Jumadi Jumadi¹, Moh Irma Sukarelawan², Heru Kuswanto¹

¹Postgraduate Program of Physics Education, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

²Postgraduate Program of Physics Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Article Info

Article history:

Received Sep 1, 2021

Revised Dec 19, 2022

Accepted Jan 3, 2023

Keywords:

Differential item functioning

Four-tier diagnostic test

Heat and temperature

Misconception

Rasch model

ABSTRACT

The existence of item bias in a set of measuring instruments can threaten the instrument's validity. Based on the Rasch model, this study evaluated item bias in the four-tier heat and temperature diagnostic test (4T-HTDT). This study used a cross-sectional quantitative survey method. There were 241 students selected using a stratified random sampling technique. The 4T-HTDT instrument consisted of 20 items grouped into five concept groups. Students' conceptual understanding was grouped into five categories, namely scientific knowledge (Rating=5), false positive (Rating=4), false negative (Rating=3), misconceptions (Rating=2), and lack of knowledge (Rating=1). The differential item functioning (DIF) score was used to evaluate item bias in the 4T-HTDT. Bias was reviewed based on the respondent's gender, class, and school. The item has DIF if the probability value is <5%. The results showed that 35% (7 out of 20 items) spread over five groups of heat and temperature concepts were biased. However, excluding seven DIF items from the measurement set would not significantly affect the composition and distribution of items. Thus, the 13 items in the 4T-HTDT instrument are free from bias and can be used to evaluate the conceptual understanding of high school students.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Moh Irma Sukarelawan

Postgraduate Program of Physics Education, Universitas Ahmad Dahlan

Umbulharjo, Yogyakarta 55161, Indonesia

Email: moh.sukarelawan@mpfis.uad.ac.id

1. INTRODUCTION

There has been an increase in the study of student's misconceptions in the field of physics education over the last seven decades, especially on the topic of heat and temperature [1]–[5]. Various types of diagnostic instruments on this topic have been widely implemented, for example, Interview [6], [7], open-ended [8], conventional multiple-choice [9], [10], two-tier diagnostic test [11], [12], three-tier diagnostic test [13]–[15], and four-tier diagnostic test [3], [11], [16], [17]. Interview and open-ended instruments can provide a more detailed and in-depth picture of students' conceptual understanding [18], [19]. However, using it in large groups takes more time and effort. As a result, interviews, and open-ended are more suitable for diagnosing misconceptions in small groups. Conventional multiple choice and their derivatives (two, three, and four-tier diagnostic test) are more suitable for large groups. The four-tier diagnostic test instrument is the best instrument for diagnostic application in large groups because the assessment process is relatively more efficient and objective, free from errors and lack of knowledge [18], [19].

One of the elements that must be met by the diagnostic test tool is validity. As a test tool, the four-tier diagnostic test in heat and temperature material (4T-HTDT) must be able to describe students' actual conceptual knowledge. The probability of students answering questions should not be influenced by student's

attributes such as region, culture, gender, school category, and other attributes [20]. If this happens, the measurement results will be biased and do not accurately reflect the situation. Bias is a condition in which the characteristics of a group affect the test results [21]. Bias also indicates injustice, inconsistency, and contaminant factors in the test instrument [20].

Measurement bias would cause assays, such as 4T-HTDT to be invariant or unequal across groups [22]. Bias causes the measurement to be irrelevant and gives certain groups a higher chance to answer correctly to benefit certain groups. This will threaten the validity of the items in the 4T-HTDT [23]. Consequently, detecting item bias in 4T-HTDT becomes essential.

The classical test theory (CTT) has been widely used to assess the psychometric properties of four-tier heat and temperature diagnostic instruments [3], [16], [17]. Although the CTT is a popular and well-established model, the scores derived are sample-dependent and biased towards the central score [24]. The item response theory (IRT) approach can be an effective alternative. Laliyo, Sumintono, and Panigoro [25] used a three-tier multiple choice test to measure students' conceptual understanding in chemistry. Aminudin *et al.* [26] developed an open four-tier diagnostic instrument on light-wave material. They report test reliability, item and person reliability, and separation index. Ibnu *et al.* [27] developed a four-tier instrument on Mechanics material. They focus on reporting item fit against the Rasch model, reliability, and separation index. Tumanggor *et al.* [28] developed a four-tier diagnostic test on simple harmonic motion concepts material. All authors report item fit, reliability, item difficulty index. Meanwhile, reports on using the Rasch model to assess differential item functioning in diagnostic instruments are still very limited, including diagnostic instruments with the concept of heat and temperature. Therefore, this study aimed to identify and report item bias in the 4T-HTDT through differential item functioning (DIF).

2. RESEARCH METHOD

A minimum sample size of 50 participants is required to achieve data stability at a 99% confidence level on a 1 Logit scale [29]. This survey involved 241 participants who were selected using a stratified random sampling technique. All respondents came from three public high schools in Indonesia. The three schools each represented schools in the favorite, moderately favorite, and less favorite categories. The age of the respondents was in the range of 15-19 years. Participation in this study was anonymous, and responses were entirely voluntary. Table 1 describes the demographic profile of students.

Table 1. Demographic data of students

Demographics (code)		Frequency	Percentage (%)
Gender	Male (M)	65	27%
	Female (F)	176	73%
Class	11 (S)	146	61%
	12 (T)	95	39%
School	Favorite (P)	113	47%
	Moderately favorite (Q)	61	25%
	Less favorite (R)	67	28%
Age (years)	Range	15-19	
	Average	17	

This study used a cross-sectional quantitative survey method. The 4T-HTDT instrument consisted of 20 items and was grouped into five concept groups, namely temperature (six items, A1-A6), thermal expansion (four items, B1-B4), the effect of heat on object temperature (two items, C1 and C2), the effect of heat on phase changes (two items, D1 and D2), and heat and heat transfer (six items, E1-E6). As the name implies, each item in the 4T-HTDT is a four-level question. The first tier (T1) is a conceptual question in the form of multiple choice. The third tier (T3) is the reasoning for responses at T1. The second (T2) and fourth (T4) tiers are each student's confidence in the responses at T1 and T3. The level of confidence (T2 and T4) uses a 6-point Likert scale from 1 (guessing only) to 6 (very confident) [30].

Students collected responses online. Completion of the diagnostic test took about 40 to 50 minutes. Students' conceptual understanding was coded using Excel. Winsteps was used to assess the 4T-HTDT instrument based on the Rasch model. Code 1 was given for correct answers on T1 and T3, zero if wrong. While code 1 was given to T2 and T4 if the student's confidence level is >3.5 , zero if <3.5 . Scientific knowledge if all tiers were coded 1. False positive if all tiers were coded 1 except T3, and false negative if all tiers were coded 1 except T1. meanwhile, misconception if T1 and T3 are coded 0, and T2 and T4 are coded 1. Other combinations of codes were included in the category of lack of knowledge.

The combination of the results of coding students' answers was grouped into five categories of conceptual understanding, namely scientific knowledge (SK, Rating=5), false positive (FP, Rating=4), false negative (FN, Rating=3), misconception (Misc, Rating=2), and lack of knowledge (LK, Rating=1) [31], [32]. Ratings of each conceptual understanding have been used to measure the Rasch model. This ordinal data was converted into intervals on a logit scale using a logarithmic function with the help of Winsteps [33], [34]. Winsteps analysis results were used to assess the logit value of the item (LVI), logit value of a person (LVP), and 4T-HTDT bias towards demographic profiles (gender, class, and school). Bias was evaluated based on the value of DIF. Meanwhile, the logit item and person values are assessed based on LVI and LVP [35], [36]. Items in the 4T-HTDT were biased towards a particular demographic profile if the probability value was less than 5% [37], [38]. Summary statistics for 4T-HTDT are presented in Table 2.

Table 2. Summary of 4T-HTDT statistics

		Item	Person
Measure	Mean	0.00	-1.33
	Standard deviation (SD)	0.41	1.83
	Separation	5.23	2.25
	Reliability	0.96	0.84
	Cronbach's α	0.92	

3. RESULTS AND DISCUSSION

3.1. Item difficulty level

The researchers classified the difficulty of the items as presented in Table 3 according to their mean and SD values in the 4T-HTDT. The difficulty levels of the items were grouped into four categories [35], [39], [40]. Item difficulty logit ranged from -0.92 to +0.58. The mean value item logit obtained was 0.00, and SD was 0.41 logit. The mean logit item value was always set to 0.0 logit, and 0.0 indicates the initial reference point of the scale [37]. Items in the 4T-HTDT were distributed in the categories very difficult ($LVI \geq 0.41$), difficult ($0.00 \leq LVI < 0.41$), easy ($-0.41 \leq LVI < 0.00$), and very easy ($LVI < -0.41$). Based on Table 3, 25% (5 out of 20 items) are in the very difficult category, and 10% (2 out of 20 items) are in the very easy category. Most items have been spread in the easy category, amounting to 35%.

As shown in Table 3, thermal expansion and heat and heat transfer concepts tend to be more difficult for students than other concepts. As many as three of the four Thermal expansion concept items are in the difficult and very difficult categories. Likewise, five heat and heat transfer concepts items are in the difficult and very difficult categories. Meanwhile, in the temperature concept, (4 out of 6 items) are divided into easy and very easy categories; two items, in the effect of heat on phase changes concept, are in the easy category. However, the distribution of each item in each concept group is represented in each level of difficulty. This has implications for the instrument's ability to explore respondents' abilities. The level of difficulty of this question needs to be considered by the teacher or instructor. Misconceptions are very vulnerable to change [3]. Because they believe wrong concepts to be true, misconceptions need to be overcome through various appropriate learning designs. Teachers need to strengthen concept planting through various innovative learning strategies. Optimizing the integration of multimedia in teaching needs to be considered by the teacher. Because of its abstract nature, changes in physical phenomena occur at the microscopic level and are theoretical, the use of simulation or animation media in heat and temperature learning is an option that needs to be considered [13], [41]–[44].

Table 3. Four item difficulty categories in 4T-HTDT

Difficulty level	Concept group				
	A	B	C	D	E
Very difficult	A2	B2	C1		E1, E2
Difficult	A1	B3, B4			E3, E4, E5
Easy	A3, A5, A6	B1	C2	D1, D2	
Very easy	A4				E6

Note: A=Temperature; B=Thermal expansion; C=Effect of heat on object temperature; D=Effect of heat on phase changes; E=Heat and heat transfer

3.2. Person ability

The level of student's conceptual understanding of heat and temperature material was grouped based on demographics and LVP values [35], [39], [40]. Student logit ability is between -5.48 and +0.63. Students' conceptual understandings are grouped into four levels. Students who have an LVP ≥ 0.50 are classified as

very high ability, first level. In the second level, the value of $-1.33 \leq LVP < 0.50$ is classified as high ability. Students with $-3.16 \leq LVP < -1.33$ are grouped in the moderate ability, third level. The fourth level, the low ability group, is used to classify students with $LVP < -3.16$. Table 4 describes the level of students' conceptual understandings based on demographics. As shown in Table 4, most students' conceptual understandings are distributed in the high category.

Table 4. The level of students' conceptual understanding based on demographics

Demographics		Very high ability $LVP \geq 0.50$	High ability $-1.33 \leq LVP < 0.50$	Moderate ability $-3.16 \leq LVP < -1.33$	Low ability $LVP < -3.16$
Gender	Male (M)	2	51	8	4
	Female (F)	1	110	33	32
Class	11 (S)	2	102	22	20
	12 (T)	1	59	19	16
School	Favorite (P)	1	78	16	18
	Moderately favorite (Q)	2	37	13	9
	Less favorite (R)	-	46	12	9

Table 4 shows that the distribution of students' conceptual understandings was concentrated in high ability. By gender, the male group appears to have a higher level of conceptual understanding than the female group. A total of 78.5% (51 out of 65 students) of male students were distributed at a high ability level. Meanwhile, only 62.5% of female students (110 out of 176 students) were distributed at the same ability level. On the other hand, the percentage of male students (6.2%) in the low ability category is much lower than that of female students (18.2%). Based on the class, 11th graders have a more dominant conceptual understanding than 12th graders. It appears that 70% (102 of 146 students) of 11th-grade students have abilities in the high category. Likewise, for grade 12 students, there are 31.2% (59 of 95 students) are in the high category. The distribution of students' abilities in these two groups did not significantly differ in the other categories (very high, moderate, and low ability).

Using a constructivism approach, gender and prior knowledge influence success in science education [45]. Male and females have different learning schemes. Various literature reports differences in conceptual understanding abilities based on gender. Male students tend to be better at observing physical phenomena than female students, which impacts their understanding of concepts [46]–[49]. However, educators or instructors need to develop pedagogic competence in designing learning to increase equality of conceptual understandings across gender. Each student's prior knowledge can be related to their age as well as their grade. The higher the students' grade level, the more sophisticated their mental models are. The construction of conceptual understanding is complete. Students at a younger age or lower grade levels can improve their conceptual understanding by involving mobile-based learning [50]–[52]. Mobile learning provides opportunities for students to access content with high frequency so that the knowledge construction process becomes faster.

Based on the type of school, the distribution of students' conceptual understanding in the three types of schools was concentrated in the high ability category. The percentage of conceptual understanding ability between favorite and less favorite schools is almost the same. The distribution of students' abilities in favorite schools was 69% (78 of 113 students), and students of less favorite schools were 68.7% (46 of 67 students). Meanwhile, the students' conceptual understanding ability at school is quite favorite by 60.7% (37 of 61 students). Meanwhile, the students' conceptual understanding ability in the low category was almost the same for the three schools. The type of school that exists does not impact the level of understanding of students' concepts. Although in general, favorite schools have several learning facilities and teachers are adequate compared to less favorite schools.

3.3. Differential item functioning (DIF) of respondents' demographic factors in 4T-HTDT

In this section, researchers identify the item bias against student demographics in public high schools. Table 5 summarizes several 4T-HTDT items with probability values less than 5%. Based on the analysis, there were seven items (A4, A6, B4, D2, E2, E5, and E6) biased towards the respondent's attributes. As many as four of the five concept groups tested were biased towards the gender, class, and type of school group. There are (4 of the 20 items) (A6, D2, E5, and E6) biased towards two types of respondents' attributes. Three (A4, B4, and E2) are biased towards one type of respondent's attributes. There are (3 of the 6 items) (E2, E5, and E6) in the heat concept group and their displacement biased towards the class and school attributes. There are (2 of the 6 items) (A4 and A6) in the temperature concept group also experienced a bias towards the class and school attributes. Meanwhile, in the concept of thermal expansion and the effect of heat on phase changes, each item (B4 and D2) is biased towards gender and school. The DIF diagram in Figure 1 describes in detail the bias tendency of each item in Table 5.

Table 5. Summary of differential item functioning based on students' demographic variables

Item	Misconception form	Demographic with DIF
A4	Two different temperatures can be added up	School
A6	The division of an object causes the temperature of the two parts to be different	Class and school
B4	An expanding substance has a constant density	Gender
D2	Heating always increases the temperature	Gender and school
E2	Cold substances contain no heat	School
E5	Heat can flow due to various types of substances	Class and school
E6	Particles during convection will rise to the top because the direction of the heat always goes up	Class and school

DIF analysis by gender is presented in Figure 1(a). There are 10% (2 out of 20 items) with a DIF (B4 and D2) concerning gender. The concept of B4 tends to benefit male students more than female students. On the other hand, the D2 concept benefits female students more than male students. DIF analysis by class in Figure 1(b) showed that 15% (3 of 20 items) had DIF (A6, E5, and E6). Items A6 and E5 tend to benefit grade 12 students rather than grade 11. At the same time, item E6 is the opposite, more favorable for grade 11 than grade 12. Finally, the DIF analysis by the school is shown in Figure 1(c). We found that 30% (6 of 20 items) had a DIF concerning school type. Items A4 and D2 tend to favor students from moderately favorite schools and disadvantaged students from less favorite schools. However, they are not problematic for students from favorite schools. Items A6 and E5 favored students from less favorite schools and disadvantaged students from the other two types of schools. Item E2 tends to harm students who come from moderately favorite schools. Meanwhile, item E6 benefits students from favorite and moderately favorite schools than students from less favorite schools.

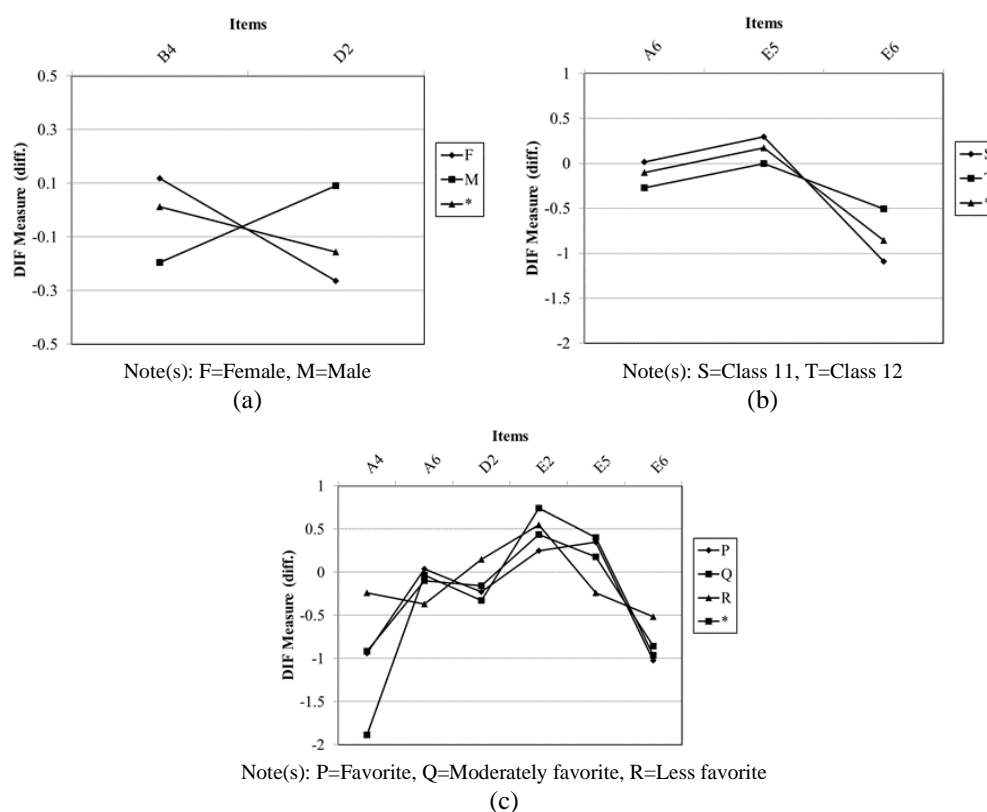


Figure 1. Person DIF plot items in 4T-HTDT based on (a) gender, (b) class, and (c) school

Seven diagnostic items on the concept of heat and temperature have been identified as being biased. When applied, these items will benefit certain groups so that the primary function of the diagnostic test is distorted and interferes with the validity of the 4T-HTDT instrument. Therefore, using 4T-HTDT in the future can exclude items identified as having DIF from the instrument set because other items still represent each concept group. Likewise, excluding seven DIF items from the measurement set will not significantly affect the distribution and composition of items based on difficulty level items.

Involving DIF analysis of items to assess the quality of a four-tier diagnostic instrument has never been reported before [3], [16], [17], [26], [28]. However, the appearance of items in the 4T-HTDT experiencing DIF does not always indicate the weakness of a measuring instrument [53]. Several studies showed the linearity of the effect of a large sample size on the number of items experiencing DIF [54].

4. CONCLUSION

One of the factors that affect the instrument's validity was the bias that the items in the instrument have. The evaluation results show that 4T-HTDT were biased. Rasch's analysis shows that seven of the 20 items (35%), spread across five heat and temperature concepts groups, are biased towards four student demographic variables (attributes). Two items are biased towards gender, three items are biased towards the type of school, and six items are biased towards the type of school.

This study still has some limitations. It has not discussed the geographical and cultural aspects of the respondents. So, this finding cannot be generalized to the Indonesian context. Nevertheless, this study has been a pioneer in evaluating item bias in the misconception diagnostic instrument. In this study, we limit the analysis of bias based on the demographics of the respondents. It took heterogeneity of respondents to evaluate instrument bias comprehensively. We suggest that future research consider the heterogeneity of the respondents who will be involved. The respondent's geography can be one of the challenging study materials to see the tendency or bias of an instrument. Schools and student residences are very diverse, some of them scattered in urban areas, some in coastal areas, and some in mountainous areas. We also recommend that a bias evaluation be carried out by reviewing the ethnicity or culture of the respondents.

ACKNOWLEDGEMENTS

The authors want to send greatest gratitude to the Directorate of Research and Community Service, Ministry of Education, Culture, Research and Technology of the Republic of Indonesia for granting research funds number: T/11.6/UN34.21/PT.01.03/2021 in the doctoral dissertation scheme.

REFERENCES




- [1] I. Aykutlu, S. Bezen, and C. Bayrak, "An assessment of high school students' conceptual structures of heat and temperature through concept maps," in *AIP Conference Proceedings*, 2017, vol. 1815, p. 070002, doi: 10.1063/1.4976423.
- [2] A. Eryilmaz, "Development and application of three-tier heat and temperature test: Sample of bachelor and students graduate," *Egitim Arastirmalari-Eurasian Journal of Educational Research*, no. 40, pp. 53–76, 2010, [Online]. Available: https://ejer.com.tr/wp-content/uploads/2021/01/ejer_2010_issue_40.pdf#page=54.
- [3] K. Fenditasari, Jumadi, E. Istiyono, and Hendra, "Identification of misconceptions on heat and temperature among physics education students using four-tier diagnostic test," *Journal of Physics: Conference Series*, vol. 1470, no. 1, p. 012055, 2020, doi: 10.1088/1742-6596/1470/1/012055.
- [4] M. R. Luce and M. A. Callanan, "Family conversations about heat and temperature: Implications for children's learning," *Frontiers in Psychology*, vol. 11, 2020, doi: 10.3389/fpsyg.2020.01718.
- [5] H. E. Haryono, K. N. Aini, A. Samsudin, and P. Siahaan, "Reducing the students' misconceptions on the theory of heat through cognitive conflict instruction (CCI)," in *AIP Conference Proceedings*, 2021, vol. 2330, p. 050001, doi: 10.1063/5.0043400.
- [6] S. H. Paik, B. K. Cho, and Y. M. Go, "Korean 4-To 11-year-old student conceptions of heat and temperature," *Journal of Research in Science Teaching*, vol. 44, no. 2, pp. 284–302, 2007, doi: 10.1002/tea.20174.
- [7] D. Ratnasari, Sukarmin, and S. Suparmi, "Effect of problem type toward students' conceptual understanding level on heat and temperature," *Journal of Physics: Conference Series*, vol. 909, no. 1, p. 012054, 2017, doi: 10.1088/1742-6596/909/1/012054.
- [8] H. Celik, "An examination of cross sectional change in student's metaphorical perceptions towards heat, temperature and energy concepts," *International Journal of Education in Mathematics, Science and Technology*, vol. 4, no. 3, p. 229, 2016, doi: 10.18404/ijemst.86044.
- [9] H. Chu, D. Treagust, S. Yeo, and M. Zadnik, "Evaluation of students' understanding of thermal concepts in everyday contexts," *International Journal of Science Education*, vol. 34, no. 10, pp. 1509–1534, 2012, doi: 10.1080/09500693.2012.657714.
- [10] M. Prince, M. Vigeant, and K. Nottis, "Development of the heat and energy concept inventory: Preliminary results on the prevalence and persistence of engineering students' misconceptions," *Journal of Engineering Education*, vol. 101, no. 3, pp. 412–438, 2012, doi: 10.1002/j.2168-9830.2012.tb00056.x.
- [11] M. I. Sukarelawan, S. Sriyanto, A. D. Puspitasari, D. Sulisworo, and U. N. Hikmah, "Four-tier heat and temperature diagnostic test (4T-HTDT) to identify student misconceptions," *JIPFRI (Jurnal Inovasi Pendidikan Fisika dan Riset Ilmiah)*, vol. 5, no. 1, pp. 1–8, 2021, doi: 10.30599/jipfri.v5i1.856.
- [12] N. Maunah and Wasis, "The development of two-tier multiple choice diagnostic test to analyze the learning difficulties of class X students on the material of temperature and heat," (in Indonesian), *Jurnal Inovasi Pendidikan Fisika (JIPF)*, vol. 03, no. 02, pp. 195–200, 2014, [Online]. Available: <https://jurnalmahasiswa.unesa.ac.id/index.php/5/article/view/8095>
- [13] M. L. H. Abbas, "Development of computer based diagnostic test for student misconception on material temperature and heat," *Jurnal Pendidikan Fisika dan Keilmuan (JPfK)*, vol. 6, no. 1, p. 12, 2020, doi: 10.25273/jpfk.v6i1.5153.
- [14] D. Gurcay and E. Gulbas, "Development of three-tier heat, temperature and internal energy diagnostic test," *Research in Science and Technological Education*, vol. 33, no. 2, pp. 197–217, 2015, doi: 10.1080/02635143.2015.1018154.
- [15] M. I. Sukarelawan, J. Jumadi, and N. A. Rahman, "An analysis of graduate students' conceptual understanding in heat and temperature (H&T) using three-tier diagnostic test," *Indonesian Review of Physics*, vol. 2, no. 1, 2019, doi: 10.12928/irip.v2i1.910.

- [16] M. Maison, I. C. Safitri, and R. W. Wardana, "Identification of misconception of high school students on temperature and calor topic using four-tier diagnostic instrument," *Edusains*, vol. 11, no. 2, pp. 195–202, 2020, doi: 10.15408/es.v11i2.11465.
- [17] J. I. Utari and F. U. Ermawati, "Development of a Four Tier Format Misconceptions Diagnostic Test Instrument for Temperature, Heat, and Displacement Materials," (in Indonesian), *Inovasi Pendidikan Fisika*, vol. 07, no. 03, pp. 434–439, 2018.
- [18] Soeharto, B. Csapó, E. Sarimanah, F. I. Dewi, and T. Sabri, "A review of students' common misconceptions in science and their diagnostic assessment tools," *Jurnal Pendidikan IPA Indonesia*, vol. 8, no. 2, pp. 247–266, 2019, doi: 10.15294/jpii.v8i2.18649.
- [19] D. Kaltakci-Gurel, A. Eryilmaz, and L. C. McDermott, "A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science," *Eurasia Journal of Mathematics, Science & Technology Education*, vol. 11, no. 5, pp. 989–1008, 2015, doi: 10.12973/eurasia.2015.1369a.
- [20] M. Ardiyaningrum, L. Badriah, Trisniawati, Suhartini, and S. A. Widodo, "Differential item function of gender in the mathematics elementary school tryout test," *Journal of Physics: Conference Series*, vol. 1315, no. 1, p. 012036, 2019, doi: 10.1088/1742-6596/1315/1/012036.
- [21] R. Akcan and K. Atalay Kabasakal, "An Investigation of Item Bias of English Test: The Case of 2016 Year Undergraduate Placement Exam in Turkey," *International Journal of Assessment Tools in Education*, vol. 6, no. 1, pp. 48–62, 2019, doi: 10.21449/ijate.508581.
- [22] K. Sadeghi and Z. Abolfazli Khonbi, "An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory," *Language Testing in Asia*, vol. 7, no. 1, p. 7, 2017, doi: 10.1186/s40468-017-0038-z.
- [23] N. D. Myers, E. W. Wolfe, D. L. Feltz, and R. D. Penfield, "Identifying differential item functioning of rating scale items with the rasch model: An introduction and an application," *Measurement in Physical Education and Exercise Science*, vol. 10, no. 4, pp. 215–240, 2006, doi: 10.1207/s15327841mpeel1004_1.
- [24] K. D. Bradley, M. R. Peabody, K. S. Akers, and N. M. Knutson, "Rating scales in survey research: using the Rasch model to illustrate the middle category measurement flaw," *Survey Practice*, vol. 8, no. 1, pp. 1–12, 2015, doi: 10.29115/sp-2015-0001.
- [25] L. A. R. Laliyo, B. Sumintono, and C. Panigoro, "Measuring changes in hydrolysis concept of students taught by inquiry model: stacking and racking analysis techniques in Rasch model," *Heliyon*, vol. 8, no. 3, Mar. 2022, doi: 10.1016/j.heliyon.2022.e09126.
- [26] A. H. Aminudin, R. Adimayuda, I. Kaniawati, E. Suhendi, A. Samsudin, and B. Coştu, "Rasch analysis of multitier open-ended light-wave instrument (MOLWI): Developing and assessing second-years Sundanese-scholars alternative conceptions," *Journal for the Education of Gifted Young Scientists*, vol. 7, no. 3, pp. 557–579, Sep. 2019, doi: 10.17478/jegys.574524.
- [27] M. Ibnu, B. Indriyani, H. Inayatullah, and Y. Guntara, "Application of the Rasch Model: Development of a Test Instrument to Measure Student Misconceptions," (in Indonesian) *Prosiding Seminar Nasional Pendidikan FKIP UNTIRTA*, vol. 2, no. 1, 2019, pp. 205–210, [Online]. Available: <https://jurnal.untirta.ac.id/index.php/psnp/article/view/5669>.
- [28] A. M. R. Tumanggor, S. Supahar, E. S. Ringo, and M. D. Harliadi, "Detecting students' misconception in simple harmonic motion concepts using four-tier diagnostic test instruments," *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, vol. 9, no. 1, pp. 21–31, 2020, doi: 10.24042/jipfalbiruni.v9i1.4571.
- [29] C. Jong, T. E. Hodges, K. D. Royal, and R. Welder, "Instruments to Measure Elementary Preservice Teachers' Conceptions: An Application of the Rasch Rating Scale Model," *Educational Research Quarterly*, vol. 39, no. 1, pp. 21–48, 2015.
- [30] B. Sreenivasulu and R. Subramaniam, "Exploring undergraduates' understanding of transition metals chemistry with the use of cognitive and confidence measures," *Research in Science Education*, vol. 44, no. 6, pp. 801–828, 2014, doi: 10.1007/s11165-014-9400-7.
- [31] Habiddin and E. M. Page, "Development and validation of a four-tier diagnostic instrument for chemical kinetics (FTDICK)," *Indonesian Journal of Chemistry*, vol. 19, no. 3, pp. 720–736, 2019, doi: 10.22146/ijc.39218.
- [32] L. A. R. Laliyo, S. Hamdi, M. Pikoli, R. Abdullah, and C. Panigoro, "Implementation of Four-Tier Multiple-Choice Instruments Based on the Partial Credit Model in Evaluating Students' Learning Progress," *European Journal of Educational Research*, vol. 10, no. 2, pp. 825–840, 2021, doi: 10.12973/eu-jer.10.2.825.
- [33] D. Adams, M. T. H. Joo, B. Sumintono, and O. S. Pei, "Blended Learning Engagement in Public and Private Higher Education Institutions: A Differential Item Functioning Analysis of Students' Backgrounds," *Malaysian Journal of Learning and Instruction*, vol. 17, no. 1, pp. 133–158, 2020, doi: 10.32890/mjli2020.17.1.6.
- [34] W. J. Boone, J. S. Townsend, and J. R. Staver, "Utilizing multifaceted Rasch Measurement through FACETS to evaluate science education data sets composed of judges, respondents, and rating scale items: An exemplar utilizing the elementary science teaching analysis matrix instrument," *Science Education*, vol. 100, no. 2, pp. 221–238, 2016, doi: 10.1002/sce.21210.
- [35] S. L. Rusland, N. I. Jaafar, and B. Sumintono, "Evaluating knowledge creation processes in the Royal Malaysian Navy (RMN) fleet: Personnel conceptualization, participation and differences," *Cogent Business and Management*, vol. 7, no. 1, 2020, doi: 10.1080/23311975.2020.1785106.
- [36] B. Setiawan, M. Panduwangi, and B. Sumintono, "A Rasch analysis of the community's preference for different attributes of Islamic banks in Indonesia," *International Journal of Social Economics*, vol. 45, no. 12, pp. 1647–1662, 2018, doi: 10.1108/IJSE-07-2017-0294.
- [37] B. Sumintono and W. Widhiarso, *Rasch Modeling Applications in Educational Assessment*. Cimahi: Trim Komunikata (in Indonesian), 2015.
- [38] M. Müller and A. Haenni Hoti, "Item analysis of the KIDSCREEN-10 using Rasch modelling," *Health and Quality of Life Outcomes*, vol. 18, no. 1, p. 342, 2020, doi: 10.1186/s12955-020-01596-6.
- [39] Maryati, Z. K. Prasetyo, I. Wilujeng, and B. Sumintono, "Measuring teachers' pedagogical content knowledge using many-facet Rasch model," *Cakrawala Pendidikan*, vol. 38, no. 3, pp. 452–464, 2019, doi: 10.21831/cp.v38i3.26598.
- [40] D. Adams, K. M. Chuah, B. Sumintono, and A. Mohamed, "Students' readiness for e-learning during the COVID-19 pandemic in a South-East Asian university: a Rasch analysis," *Asian Education and Development Studies*, vol. 11, no. 2, pp. 324–339, 2022, doi: 10.1108/AEDS-05-2020-0100.
- [41] I. Kaniawati, G. Triyani, A. Danawan, I. Suyana, A. Samsudin, and E. Suhendi, "Implementation of Interactive Conceptual Instruction (ICI) With Computer Simulation: Impact of Students' Misconceptions on Momentum and Impulse Material," *Jurnal Ilmiah Pendidikan Fisika Al-BiRuNi*, vol. 10, no. 1, pp. 1–17, 2021, doi: 10.24042/jipfalbiruni.v10i1.8375.
- [42] C. Chou, "An Analysis of the 3D Video and Interactive Response Approach Effects on the Science Remedial Teaching for Fourth Grade Underachieving Students," *Journal of Mathematics Science and Technology Education*, vol. 13, no. 4, pp. 1059–1073, 2017, doi: 10.12973/eurasia.2017.00658a.
- [43] Y. Kong, L. R. Kayumova, and V. G. Zakirova, "Simulation technologies in preparing teachers to deal with risks," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 8, pp. 4753–4763, 2017, doi: 10.12973/eurasia.2017.00962a.




- [44] E. D. Sanyoto, W. Setyarsih, and A. Kholiq, "Application of Interactive Demonstration Learning Models Assisted by Virtual Simulation Media to Reduce Students' Misconceptions on Temperature, Heat, and Heat Transfer Materials," (in Indonesian), *Jurnal Inovasi Pendidikan Fisika (JIPF)*, vol. 05, no. 03, pp. 188–192, 2016.
- [45] A. Taşdere and F. Ercan, "An alternative method in identifying misconceptions: Structured communication grid," in *Procedia - Social and Behavioral Sciences*, vol. 15, pp. 2699–2703, 2011, doi: 10.1016/j.sbspro.2011.04.173.
- [46] R. Scherr, "Editorial: Never mind the gap: gender-related research in physical review physics education research, 2005-2016," *Physical Review Physics Education Research*, vol. 12, no. 2, p. 020003, 2016, doi: 10.1103/PhysRevPhysEducRes.12.020003.
- [47] R. Koul, T. Lerdpornkulrat, and C. Poondej, "Gender compatibility, math-gender stereotypes, and self-concepts in math and physics," *Physical Review Physics Education Research*, vol. 12, no. 2, p. 020115, 2016, doi: 10.1103/PhysRevPhysEducRes.12.020115.
- [48] A. Madsen, S. B. McKagan, and E. C. Sayre, "Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?" *Physical Review Special Topics - Physics Education Research*, vol. 9, no. 2, p. 020121, 2013, doi: 10.1103/PhysRevSTPER.9.020121.
- [49] N. Munfarikha, S. Kusairi, and S. Zulaikhah, "Effect of IQ, gender and grade level on high school students' mental models of magnetism," (in Indonesian), *QUANTUM, Seminar Nasional Fisika dan Pendidikan Fisika*, 2018, pp. 487–494.
- [50] B. E. Dasilva *et al.*, "Development of android-based interactive physics mobile learning media (IPMLM) with scaffolding learning approach to improve HOTS of high school students," *Journal for the Education of Gifted Young Scientists*, vol. 7, no. 3, pp. 659–681, 2019, doi: 10.17478/jegys.610377.
- [51] F. P. Sari, L. Ratnaningtyas, I. Wilujeng, Jumadi, and H. Kuswanto, "Development of android comics media on thermodynamic experiment to map the science process skill for senior high school," *Journal of Physics: Conference Series*, vol. 1233, no. 1, p. 012052, 2019, doi: 10.1088/1742-6596/1233/1/012052.
- [52] F. S. Arista and H. Kuswanto, "Virtual physics laboratory application based on the android smartphone to improve learning independence and conceptual understanding," *International Journal of Instruction*, vol. 11, no. 1, pp. 1–16, 2018, doi: 10.12973/iji.2018.1111a.
- [53] P. Susongko, Y. Arfiani, and M. Kusuma, "Determination of gender differential item functioning in Tegal students' scientific literacy skills with integrated science (SLiSIS) test using Rasch model," *Jurnal Pendidikan IPA Indonesia*, vol. 10, no. 2, pp. 270–281, 2021, doi: 10.15294/jpii.v10i2.26775.
- [54] R. Zwick, "a review of Ets differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement," *ETS Research Report Series*, vol. 2012, no. 1, pp. i–30, 2012, doi: 10.1002/j.2333-8504.2012.tb02290.x.

BIOGRAPHIES OF AUTHORS






Jumadi    is a Professor in the Physics Education Study Program and Postgraduate Program at Universitas Negeri Yogyakarta, Indonesia. He received his Doctor of Science Education degree in 2002 from the Universitas Pendidikan Indonesia. His area of expertise is in Physics Learning Technology. Currently, his research focus is on the development of science learning in the 21st century based on a scientific approach to improve scientific literacy. He can be contacted at email: jumadi@uny.ac.id.



Moh Irma Sukarelawan    is a Ph.D. Candidate, Department of Educational Science, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta 55281, Indonesia & Lecturer, Postgraduate Program of Physics Education, Universitas Ahmad Dahlan, Indonesia. His research focuses on physics education, misconception, and Rasch model. He can be contacted at email: moh.irma2016@student.uny.ac.id; moh.sukarelawan@mpfis.uad.ac.id.



Heru Kuswanto    is a professor at the Physics Education Study Program and Postgraduate Program at Universitas Negeri Yogyakarta, Indonesia. He received his Doctorate in Physics in 2002 from Jean Monnet De Saint Etienne, France. His areas of expertise are Optics, Optoelectronics, and Microwaves. He can be contacted at email: herukus61@uny.ac.id.